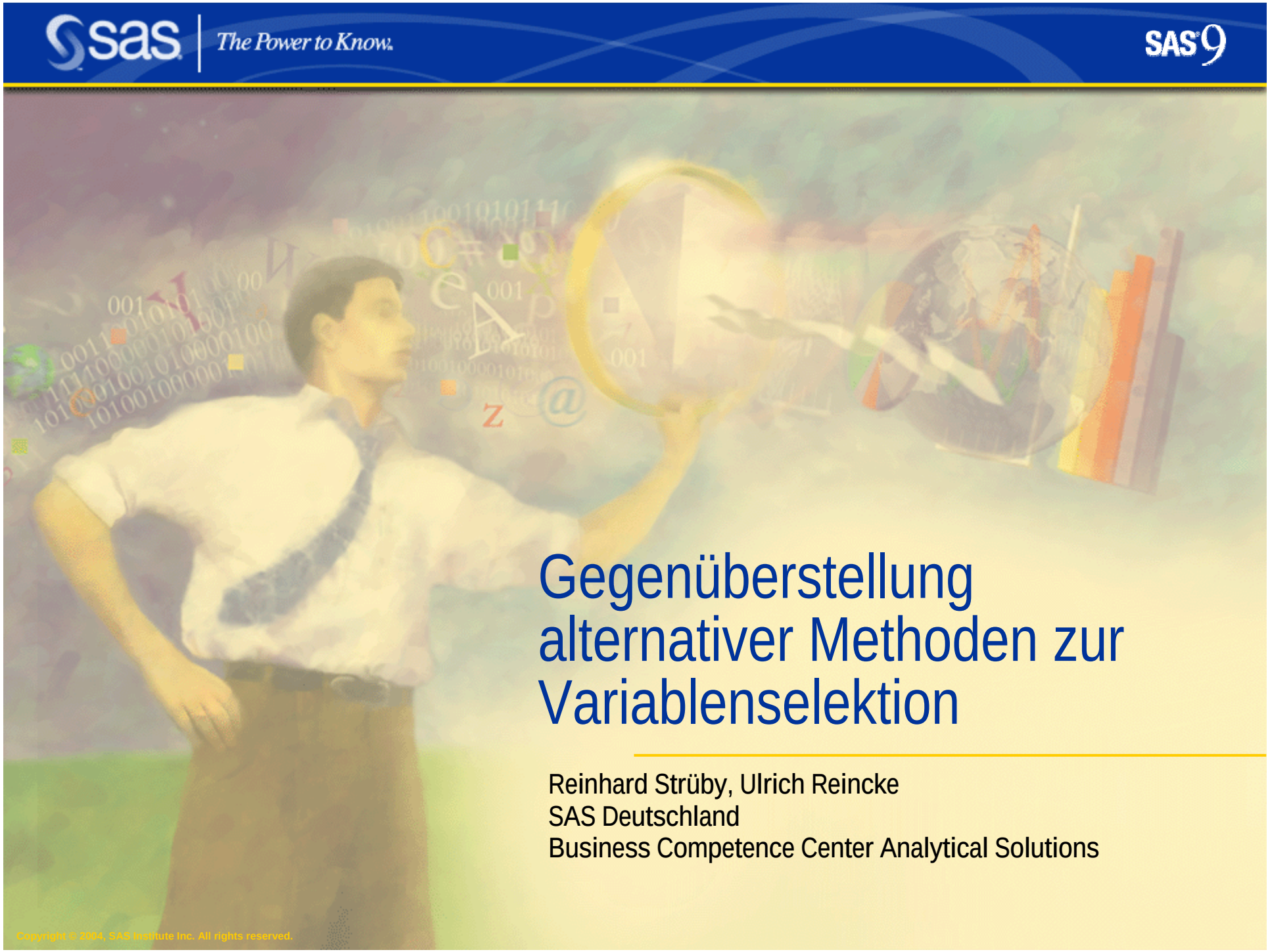


Gegenüberstellung alternativer Methoden zur Variablenselektion

Reinhard Strüby, Ulrich Reincke
SAS Deutschland
Business Competence Center Analytical Solutions

Klassische Modellwahl mit SAS

PROC REG data=indata;

MODEL depvar=var1--var200 / selection=forward;

RUN;

PROC LOGISTIC data=indata;

MODEL depvar=var1--var200 / selection=stepwise;

RUN;

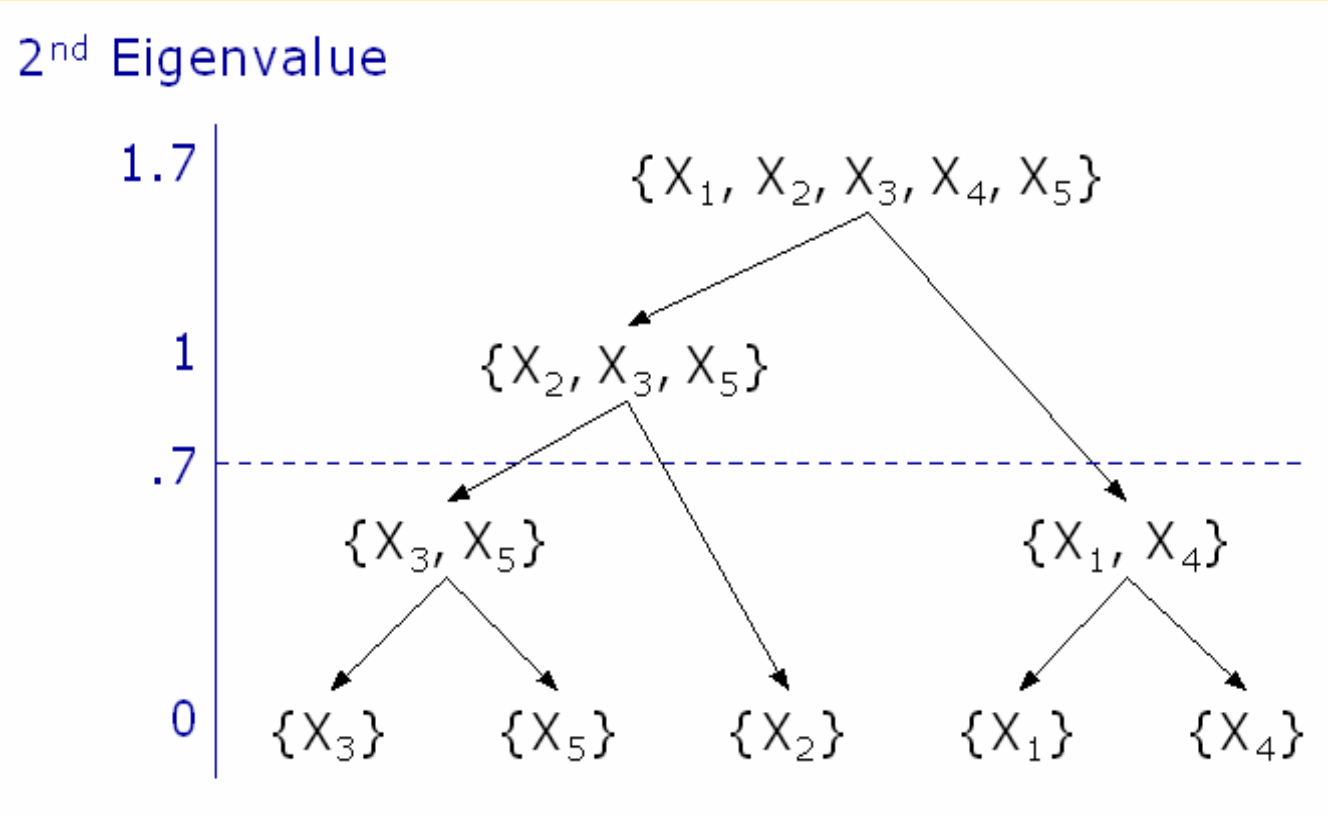
Die Prozedur STEPDISC

```
PROC STEPDISC DATA = data-set METHOD = method;
     CLASS variable;
     VAR variables;
RUN;
```

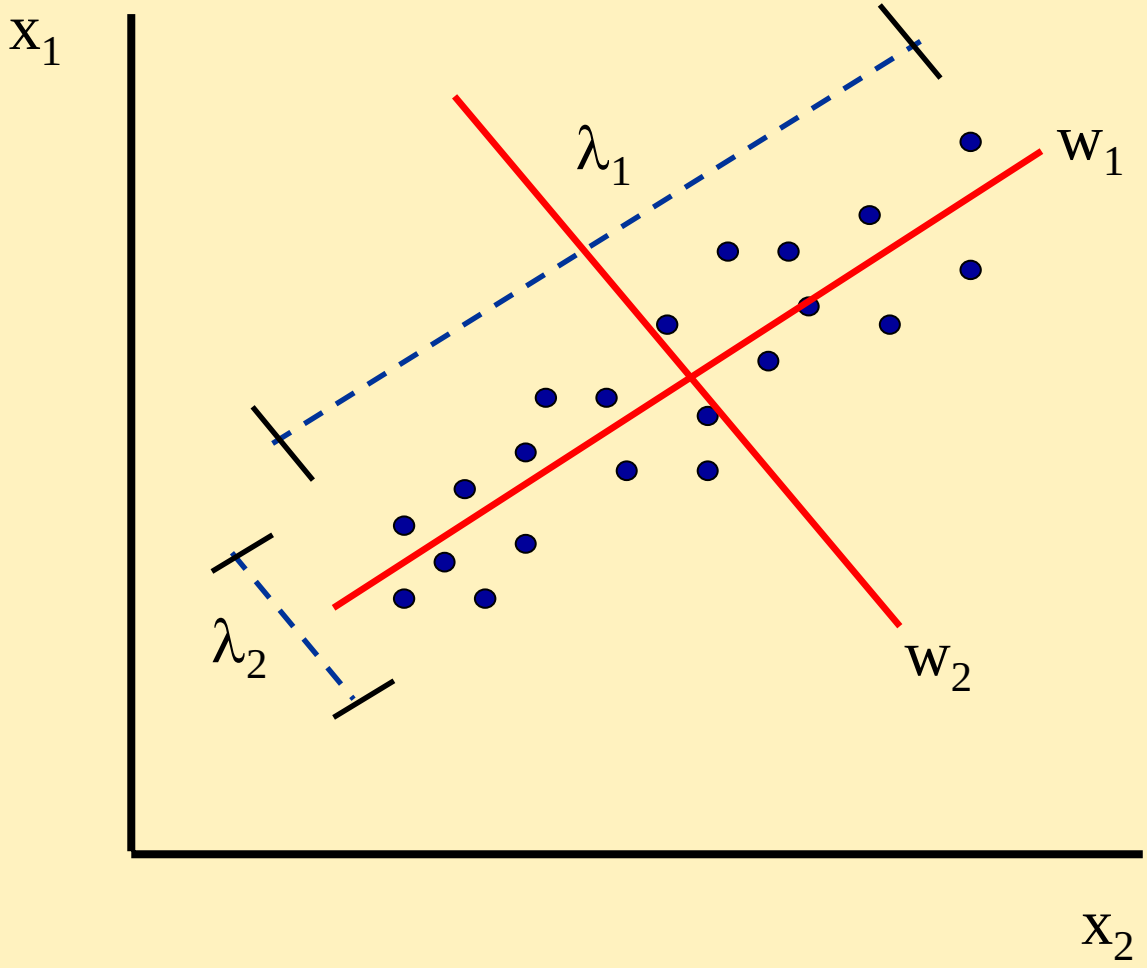
- Modellwahl über kanonische Variable
- Beurteilung der Klassifikationsgüte mit PROC DISCRIM

Divisive Clustering

- PROC VARCLUS nutzt Divisive Clustering um Untergruppen von Variablen zu bilden, die sich möglichst stark unterscheiden.



Eigenwerte und Eigenvektoren



Variablen-Reduktion

Cluster	Variable	R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio
Cluster 1	RedMeat	0.5350	0.2185	0.5950
	WhiteMeat	0.4544	0.3331	0.8181
	Eggs	0.7926	0.4902	0.4067
	Milk	0.5529	0.2721	0.6142
Cluster 2	Cereal	0.8255	0.4630	0.3250
	Nuts	0.8255	0.4549	0.3201
Cluster 3	Fish	0.7019	0.1365	0.3452
	Starch	0.7019	0.3075	0.4304
Cluster 4	FruitVeg	1.0000	0.0538	0.0000

- PROC VARCLUS reduziert Variable, nicht Dimensionen.

Neue Prozedur GLMSELECT Schwerpunkt Variablenselektion

- GLMSELECT bildet Lineare Modelle ähnlich REG oder GLM.
- Schwerpunkt auf der Modellwahl !
  - Keine Regressionsdiagnostiken
  - Keine Hypothesentests
  - Keine Kontraste
  - Keine LS-Means Analysen

Modernste Modellwahlverfahren

- LASSO Methode nach Tibshirani (1996)
  - Analog OLS
  - Die Summe der Beträge der Parameterschätzungen wird beschränkt.
  - CLASS-Variable werden aufgesplittet.
- LAR Methode nach Efron et. Al (2004)
  - Least Angle Regression
  - Gegenüber LSE „geschrumpfte“ Parameter
  - CLASS-Variable werden aufgesplittet.

Wichtige Eigenschaften der Prozedur GLMSELECT

- Vielzahl von Selektionskriterien CHOOSE=
  - ADJRSQ Adjusted R-square statistic
  - AIC Akaike information criterion
  - AICC Corrected Akaike information criterion
  - BIC Sawa Bayesian information criterion
  - CP Mallow C(p) statistic
  - CV Predicted residual sum of square with k-fold cross validation
  - PRESS Predicted residual sum of squares
  - SBC Schwarz Bayesian information criterion
  - VALIDATE Average square error for the validation data
- leave-one-out und k-fold Cross Validation

Weitere Merkmale

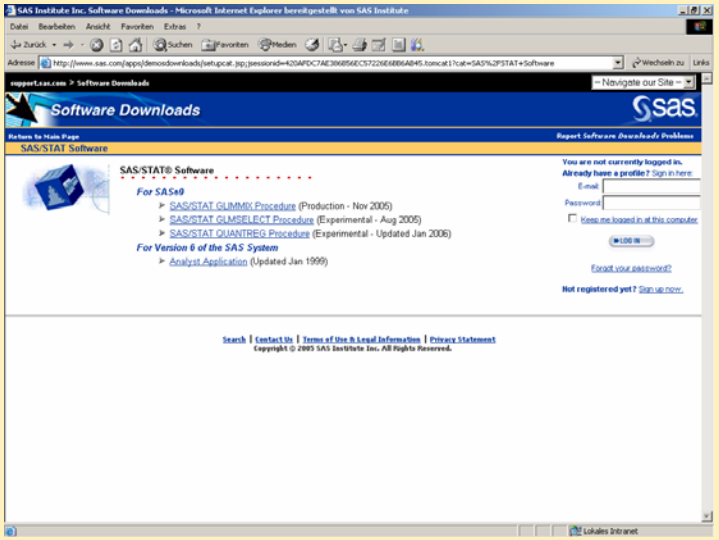
- Beliebige Interaktionen und eingebettete Effekte
- Hierarchien von Effekten
- Integriertes Partitionieren der Daten in Trainings-, Validations- und Testdatei
- Grafische Visualisierungen des Selektionsprozesses

Und schließlich

- Vorhersagen und Residuen in Output-Dateien
- Erzeugt Macro-Variable mit dem selektierten Modell
- Parallelverarbeitung von BY-Gruppen
- Große Zahl möglicher Effekte (> 10.000 kein Problem)
- SCORE Statements

Installation der experimentellen Prozedur GLMSELECT

- Download von http://support.sas.com



- Die GLMSELECT Prozedur benötigt eine SAS 9 Windows Standalone-Installation.
  1. Schließen aller SAS-Anwendungen
  2. SASGLMSELECT.exe ausführen

Syntax

PROC GLMSELECT < options >**;**
BY variables **;**
CLASS variable <(v-options)> <variable <(v-options)>... >
< / v-options > < options >**;**
FREQ variable **;**
MODEL variable = < effects >< / options >**;**
OUTPUT < **OUT=**SAS-data-set > < keyword<=name> >
< . . . keyword<=name> > **;**
PARTITION < options >**;**
PERFORMANCE < options >**;**
SCORE < **DATA=**SAS-data-set > < **OUT=**SAS-data-set > > **;**
WEIGHT variable **;**

PARTITION Statement

- **FRACTION(<TEST=fraction> <VALIDATE=fraction>)**
- Steuert bestimmte Verhältnisse für die zufällige Auswahl von Teilstichproben.
- Damit werden Trainings- und Validationsdateien generiert (analog SEMMA im SAS Data Mining).

Beispiel:

```
partition fraction(validate=0.5);
```

PERFORMANCE Statement

- **CPUCOUNT =** ACTUAL
- Gibt die Zahl der zu erwartenden physischen Prozessoren an.
- Wirksam bei BY-Gruppen-Verarbeitung.
- Überschreibt die SAS Systemoption (vorsichtig verwenden !).

Beispiel:

```
cpucount = 4
```

PERFORMANCE Statement

- **THREADS**
- Erlaubt Multithreaded-BY-Gruppenverarbeitung
- Überschreibt die SAS Systemoption THREADS | NOTHREADS.
- Ohne BY-Gruppenverarbeitung wird die Option ignoriert.

Beispiel:

```
threads
```

Score Statement

- Erzeugt neue Datei mit gescorten Daten, d.h. für jede Beobachtung werden Vorhersagen abgelegt, wahlweise auch die Residuen.
- Nähe zu operativen Anwendungen

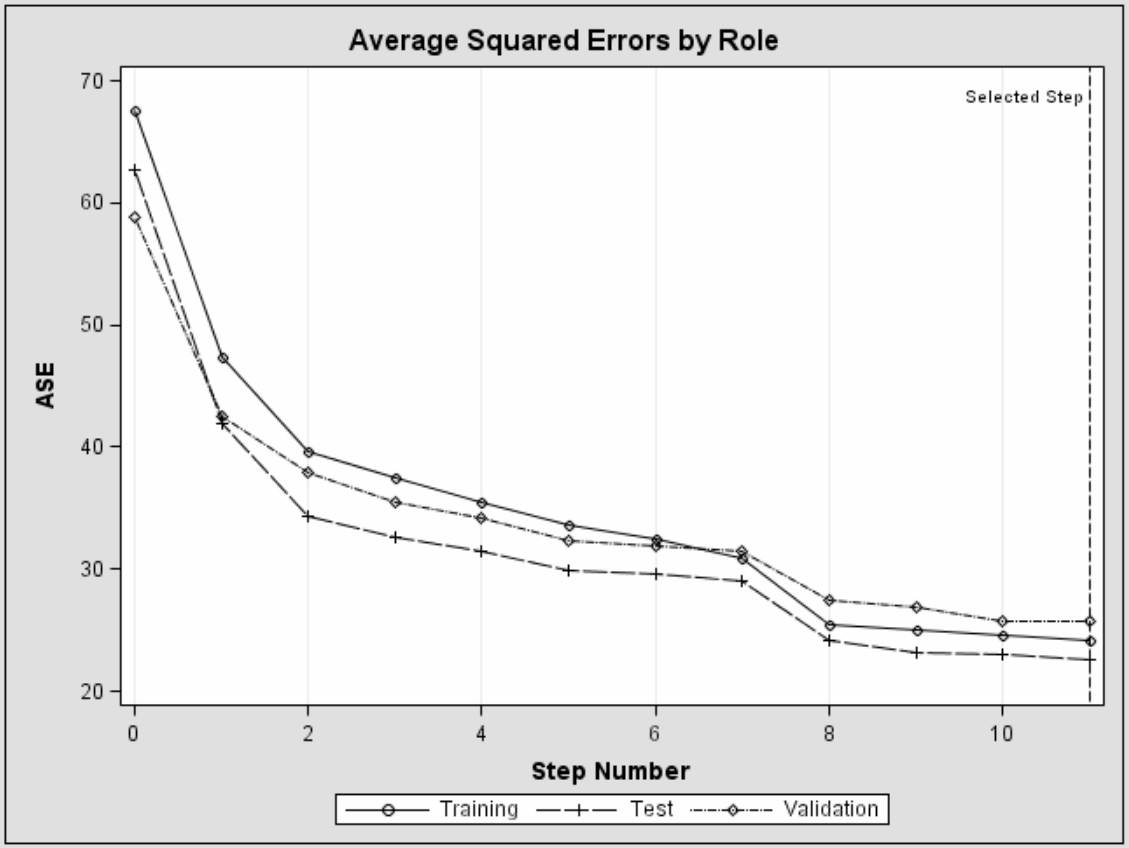
Beispiel:

```
score data=indata out=outdata pred=predvar
resid=residvar;
```

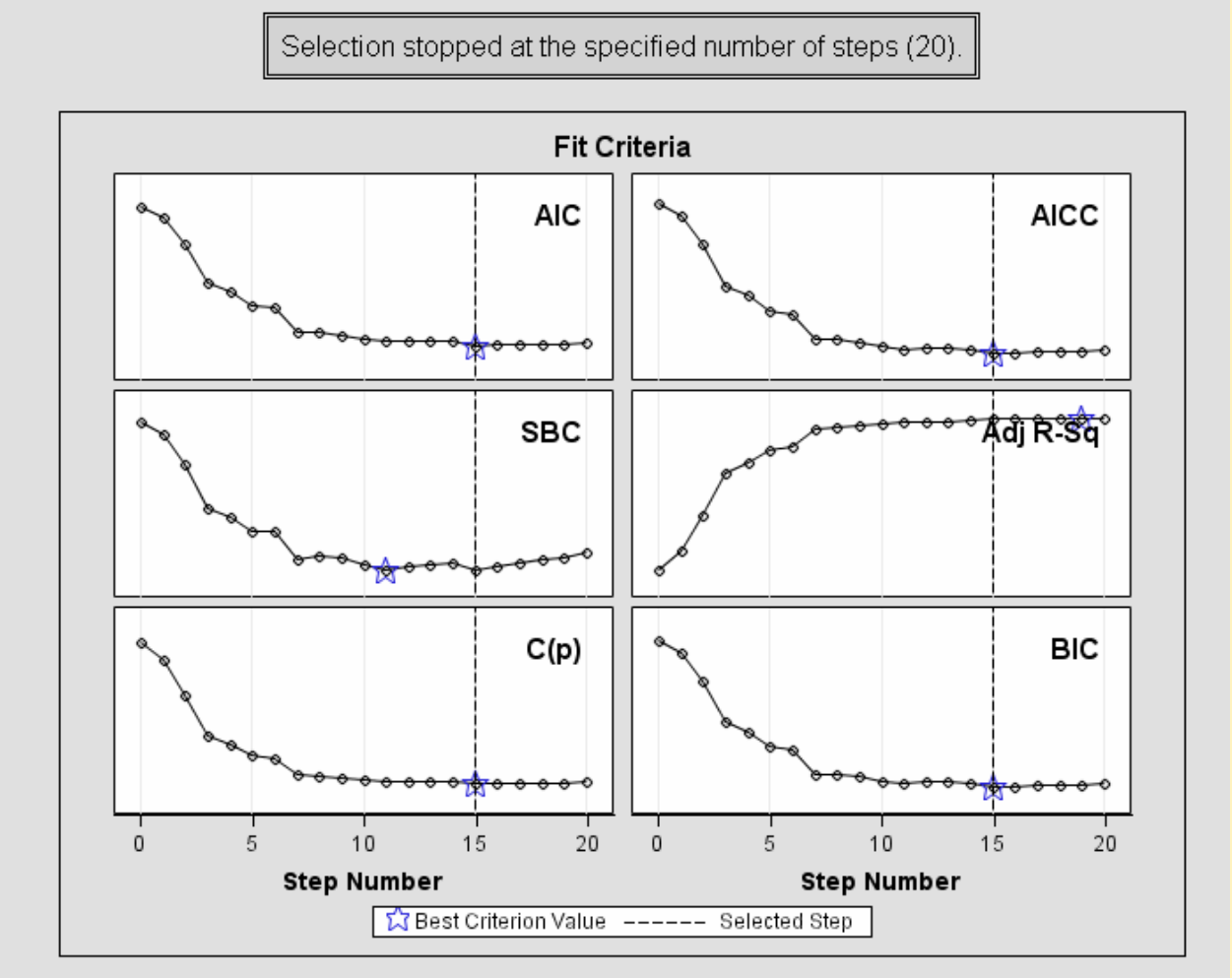
ODS Grafiken (Auswahl)

- AdjRSqPlot Adjusted R-square by step
- AICPlot Akaike Information Criterion by step
- ASEPlot Average square errors by step
- CandidatesPlot SELECT criterion by effect CANDIDATES
- ChooseCriterionPlot CHOOSE criterion by step
- CoefficientPanel Coefficients
- ...

Mittlere Quadratfehler
Train – Test - Validation



Anpassungsgüten



GLMSELECT Beispiel

```
proc glmselect data=KSFE.baseball
                    plots=(CriterionPanel ASE) seed=1;
  partition fraction(validate=0.3 test=0.2);
  class league division;
  model logSalary = nAtBat nHits nHome nRuns nRBI nBB
            yrMajor crAtBat crHits crHome crRuns crRbi
            crBB league division nOuts nAssts nError
          / selection=forward(choose=validate stop=10);
run;
```